#3 PC
1-10-00

IN THE U.S. PATENT AND TRADEMARK OFFICE

Applicant(s): NISHIMURA, Hideki

Application No.: Group:

Filed: October 19, 1999 Examiner:

For: KEY WORD DERIVING DEVICE, KEY WORD DERIVING METHOD, AND STORAGE MEDIUM CONTAINING KEY WORD DERIVING PROGRAM

L E T T E R

Assistant Commissioner for Patents
Box Patent Application
Washington, D.C. 20231

October 19, 1999
0397-0393P-SP

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55(a), the applicant hereby claims the right of priority based on the following application(s):

Country	Application No.	Filed
JAPAN	10-300720	10/22/98

A certified copy of the above-noted application(s) is(are) attached hereto.

If necessary, the Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any overpayment to deposit Account No. 02-2448 for any additional fees required under 37 C.F.R. 1.16 or under 37 C.F.R. 1.17; particularly, extension of time fees.

Respectfully submitted,

BIRCH, STEWART, KOLASCH & BIRCH, LLP

By: _____
TERRELL C. BIRCH
Reg. No. 19,382
P. O. Box 747
Falls Church, Virginia 22040-0747

Attachment
(703) 205-8000
/dll

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９８年１０月２２日

出 願 番 号
Application Number: 平成１０年特許願第３００７２０号

出 願 人
Applicant (s): シャープ株式会社

１９９９年 ８月 ２日

特 許 庁 長 官
Commissioner,
Patent Office 伊佐山建志

出証番号 出証特平１１－３０５３９２４

【書類名】　特許願

【整理番号】　98-02657

【提出日】　平成10年10月22日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/30

【発明の名称】　キーワード抽出方法、キーワード抽出装置、及びキーワード抽出プログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】　7

【発明者】

【住所又は居所】　大阪府大阪市阿倍野区長池町２２番２２号　シャープ株式会社内

【氏名】　西村　英樹

【特許出願人】

【識別番号】　000005049

【氏名又は名称】　シャープ株式会社

【電話番号】　06-621-1221

【代理人】

【識別番号】　100103296

【弁理士】

【氏名又は名称】　小池　隆彌

【電話番号】　06-621-1221

【連絡先】　電話０４３－２９９－８４６６　知的財産権センター東京知的財産権部

【手数料の表示】

【予納台帳番号】　012313

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　　図面　1

【物件名】　　要約書　1

【包括委任状番号】　9703283

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　キーワード抽出方法、キーワード抽出装置、及びキーワード抽出プログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項１】　データからキーワードを抽出する方法であって、

所望のパラメータを用いて前記データを分割するステップと、

分割されたグループ毎に単語を統計処理するステップと、

統計処理された結果を比較し、重要度を算出するステップと、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定するステップと、

を備えたことを特徴とするキーワード抽出方法。

【請求項２】　前記所望のパラメータは、前記データにそれぞれ付加される属性から選択されてなることを特徴とする請求項１に記載のキーワード抽出方法。

【請求項３】　前記分割されたグループ毎に単語を統計処理する際、全てのグループに同一の統計処理を行ってなることを特徴とする請求項１又は２に記載のキーワード抽出方法。

【請求項４】　前記キーワードは、グループ毎のキーワード、又は全データのキーワードとして決定されることを特徴とする請求項１～３のいずれかに記載のキーワード抽出方法。

【請求項５】　請求項４で決定されるキーワードは、パラメータとして選択された属性と関連づけて、これを全データの特性として判断されてなることを特徴とするキーワード抽出方法。

【請求項６】　データからキーワードを抽出する装置であって、

所望のパラメータを用いて前記データを分割する手段と、

分割されたグループ毎に単語を統計処理する手段と、

統計処理された結果を比較し、重要度を算出する手段と、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定する手段と、

を備えたことを特徴とするキーワード抽出装置。

【請求項７】　コンピュータを、

所望のパラメータを用いて前記データを分割する手段、

分割されたグループ毎に単語を統計処理する手段、

統計処理された結果を比較し、重要度を算出する手段、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定する手段、

として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は大量のデータから、そのデータの特徴を示すキーワードを抽出する方法に関するもので、大量のデータの部分領域の統計処理を行うことによってキーワードを抽出するキーワード抽出方法およびキーワード検索装置およびキーワード抽出プログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】

大量のデータの部分領域の統計処理を行うことによってキーワードを抽出する方法が特開平８－２０２７３７号公報に記載されている。ここでは特許の明細書を例にあげ、まず、あらかじめ準備した、《発明の名称》、《特許請求の範囲》、等の見出し語に注目して全データを個々の段落に分割し、次に、同一文（センテンス）内での各単語の他の単語との共起数、同一段落毎の各単語の他の単語との共起数、全データでの各単語の出現数を求め、最後にこれらに適当な係数を乗じた代数和で各単語の重要度を求め、キーワードを決定している。

【０００３】

すなわち、単なる各単語の出現頻度でキーワードを決定するのでは無く、同一文、同一段落で相互に共起する単語はより重要度（キーワードとしての妥当性）が高いと判断している。

【０００４】

【発明が解決しようとする課題】

しかしながら、特開平８－２０２７３７号公報に記載の方法では、対象データの特殊性に基づいて予め準備された見出し語（《発明の名称》等）によって段落分割がなされるため、段落の分割が固定であった。また、抽出されるキーワードは対象データ全体に対するキーワードであって、個々の段落のキーワードは抽出されていなかった。

【０００５】

従って、対象データが、特許明細書のように各段落がそれぞれ固定の意味を持ち、明細書１文書で内容が完結しているような場合には問題が少なかったが、対象データが各個人が送受信した（電子）メール全体とか、１日、１月単位のニュース全体とかのように、送受信相手、発生時刻（日時）、等の種々のパラメータで分割可能なデータ集合であり、対象データ全体の内容が把握しにくい場合には適用できなかった。

【０００６】

本発明は、上記課題に基づいて創案されたもので、種々のパラメータで分割可能な大量のデータに対して各分割毎の統計処理結果の違いを比較することにより、各分割毎または全データのキーワードを抽出し、全データの特徴と、全データの中で各分割の特異性、傾向との両方または一方を把握することを目的とする。

【０００７】

【課題を解決するための手段】

この発明（請求項１）に係るキーワード抽出方法は、データからキーワードを抽出する方法であって、

所望のパラメータを用いて前記データを分割するステップと、

分割されたグループ毎に単語を統計処理するステップと、

統計処理された結果を比較し、重要度を算出するステップと、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定するステップと、

を備えたことにより、上記の目的を達成する。

【０００８】

この発明（請求項２）に係るキーワード抽出方法は、請求項１において、

前記所望のパラメータが、前記データにそれぞれ付加される属性から選択されてなることにより、上記の目的を達成する。

【０００９】

この発明（請求項３）に係るキーワード抽出方法は、請求項１又は２において、

前記分割されたグループ毎に単語を統計処理する際、全てのグループに同一の統計処理を行ってなることにより、上記の目的を達成する。

【００１０】

この発明（請求項４）に係るキーワード抽出方法は、請求項１～３のいずれかにおいて、

前記キーワードが、グループ毎のキーワード、又は全データのキーワードとして決定されることにより、上記の目的を達成する。

【００１１】

この発明（請求項５）に係るキーワード抽出方法は、

請求項４で決定されるキーワードが、パラメータとして選択された属性と関連づけて、これを全データの特性として判断されてなることにより、上記の目的を達成する。

【００１２】

この発明（請求項６）に係るキーワード抽出装置は、

データからキーワードを抽出する装置であって、

所望のパラメータを用いて前記データを分割する手段と、

分割されたグループ毎に単語を統計処理する手段と、

統計処理された結果を比較し、重要度を算出する手段と、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定する手段と、

を備えたことにより、上記の目的を達成する。

【００１３】

この発明（請求項７）に係るキーワード抽出プログラムを記録したコンピュータ読み取り可能な記録媒体は、コンピュータを、

所望のパラメータを用いて前記データを分割する手段、

分割されたグループ毎に単語を統計処理する手段、

統計処理された結果を比較し、重要度を算出する手段、

算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定する手段、

として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることにより、上記の目的を達成する。

【００１４】

即ち、本発明は、大量のデータを所望のパラメータを用いて、いくつかの数に分割するものである。この所望のパラメータは、データを社外向けメールと社内向けメールに分割する、メールの発生月日または時刻で分割する、又は、特定の相手とそれ以外に分割する等、対象とするデータに付加された属性であれば、何でも良い。

【００１５】

次に、分割されたグループ毎に単語の統計処理を行う。この統計処理は単なる単語の出現頻度でも良いし、他の単語との共起数であっても良い。

【００１６】

更に、分割されたグループ毎の統計処理結果の比較を行う。比較を行う場合には、前記分割されたグループ毎の統計処理を、単なる出現頻度であれ、他の単語との共起数であれ、統一しておくことが好ましい。また、具体的な比較は統計結果の差分もしくは比率を求めることによってなされる。

【００１７】

つまり、ある分割領域において、出現頻度又は他の単語との共起数が大きく、他の分割領域において、出現頻度又は他の単語との共起数が小さい単語は、重要度（キーワードとしての妥当性）が高いと判断する。

【００１８】

最後に各分割で重要度が高いと判断された単語からキーワードを決定する。これは分割されたグループ毎のキーワードと判断しても良いし、全データのキーワードと判断しても良い。最も望ましいのは、最初に分割した時の属性と関連づけて、ある属性分割にはこのようなキーワードがあり、別の属性分割には別のキーワードがあるのが、この全データの特性であると判断することである。

【００１９】

【発明の実施の形態】

以下、本発明の一実施例を添付図面に基づいて詳細に説明する。

【００２０】

図１は本実施例に係るキーワード抽出装置の機能ブロック図である。

【００２１】

この図において、１は対象データ選択手段である。対象データ選択手段１は、対象の文書集合に対して、各文書の単語やその他付加情報を整理する。２は対象データ分割手段である。対象データ分割手段２は、各文書の付加情報を利用して、文書集合を所望のＮ個に分割する。３は部分統計処理手段である。部分統計処理手段３は、分割されたＮ個の集合それぞれに対して単語の統計をとる。４は部分統計処理結果比較手段である。部分統計処理結果比較手段４は、Ｎ個の統計を元に、統計結果を比較し、違いを検出して、単語の重要度を決定する。５はキーワード抽出手段である。キーワード抽出手段５は、単語の重要度を参照し、ソートすることで、キーワードを抽出する。

【００２２】

以後、ある開発資料の文書集合を対象に、著者を分割基準とした場合の動作を詳細に説明する。ここで、文書はコンピュータ上のファイルとなっており、著者情報を持ち、文書集合をファイル名のリストという形式で扱えるものとする。図２は、対象データ選択手段１における動作のフローチャートである。まず、文書集合から文書を１つ取り出す（Ｓ１０１）。取り出した文書を解析し、単語の出現頻度を調べる（Ｓ１０２）。この解析方法については、公知の技術を用いることにし、特に制限しない。

【００２３】

次に、不要語辞書を用いて「そして、しかし、この」などの不要語を除去した後（Ｓ１０３）、著者名とともに文書テーブルに、文書の要素として追加する（Ｓ１０４）。文書集合の全ての文書に対して行うまで、Ｓ１０１～Ｓ１０４を繰り返す（Ｓ１０５）。結果として、文書テーブル６（図３）を得る。

【００２４】

ここで、単語の後の（）の数字は、その単語の出現数を表す。なお、著者名は、ファイルの著者情報を利用するなどで、特定できるものとする。ここでは、説明を簡単にするために各文書とも単語数が極端に少ない例にしているが、実際の文書では当然のことながら単語の種類・出現頻度とも多種・多数になる。

【００２５】

図４は、対象データ分割手段２における動作のフローチャートである。文書テーブル６の全ての要素に対して、Ｓ２０１～Ｓ２０３を繰り返す。まず、文書テーブル６から、文書要素を１つ取り出す（Ｓ２０１）。例えば、「１　著者：幹部Ａ　単語：画期的（１０）、技術革新（５）、デジタル（４）」が取り出される。次に、著者名から、著者クラスを決定する（Ｓ２０２）。著者クラスとは、分割単位グループのことであり、クラスはいくつあっても良いが、この例では、「幹部」「技術」「企画」の３つのクラスがあるとする。従って、「１　著者：幹部Ａ　単語：画期的（１０）、技術革新（５）、デジタル（４）」の場合、クラス「幹部」と決定する。

【００２６】

そして、決定されたクラスに従って、該当するクラスの単語リストに単語を追加する（Ｓ２０３）。「１　著者：幹部Ａ　単語：画期的（１０）、技術革新（５）、デジタル（４）」の場合、幹部クラスの単語リスト７に、単語「画期的（１０）、技術革新（５）、デジタル（４）」を追加する（図５）。文書テーブル６の全ての要素に対して行うまで、Ｓ２０１～Ｓ２０３を繰り返す（Ｓ２０４）。

【００２７】

結果として、「幹部」「企画」「技術」の３つのクラスに対応して、単語リス

ト７（図５）、単語リスト８（図６）、単語リスト９（図７）を得る。なお、著者クラスは著者名から直接決定するとしたが、著者クラステーブル１０（図８）を用いて、クラスを決定するようにしても良い。

【００２８】

図９は、部分統計処理手段３における動作のフローチャートである。全てのクラスに対して、Ｓ３０１～Ｓ３０６を繰り返す。まず、対象データ分割手段２によって得られた、あるクラスに対する、単語リストを選択する（Ｓ３０１）。例えば、幹部クラスの単語リスト７（図５）を選択する。選択した単語リストの全ての単語に対して、Ｓ３０２～Ｓ３０５を繰り返す。選択した単語リストから、単語を１つ取り出す（Ｓ３０２）。

【００２９】

例えば、幹部クラスの単語リスト７から、単語「画期的（１０）」を取り出す。そして、取り出した単語に対して、単語カウントテーブルに登録されているかどうかを調べ（Ｓ３０３）、登録されていれば、その単語に対応するカウンタを出現数だけ増加する（Ｓ３０４）。登録されていなければ、その単語を単語カウントテーブルに登録し、カウンタを出現数に設定する（Ｓ３０５）。

【００３０】

例えば、幹部クラスを選択している時、単語リスト７から取り出された「画期的（１０）」は１回目は単語カウントテーブル１１に登録されていないので、単語カウントテーブル１１に登録し、カウンタをその出現数である１０にする。しかし、２回目に「画期的（２）」が取り出された時、単語カウントテーブル１１に登録されているので、対応するカウンタを２増加し、１２とする。

【００３１】

Ｓ３０２～Ｓ３０５を、選択した単語リストの全ての単語に対して行うまで繰り返す（Ｓ３０６）。Ｓ３０１～Ｓ３０６を、全てのクラスに対して行うまで繰り返す（Ｓ３０７）。

【００３２】

結果として、各クラスに対応した、単語カウントテーブル１１、１２、１３を得る（図１０、図１１、図１２）。

【００３３】

図１３は、部分統計処理結果比較手段４における動作のフローチャートである。Ｓ４０１～Ｓ４０４を、全てのクラスに対して行うまで繰り返す。まず、対象のクラスを選択する（Ｓ４０１）。例えば、幹部の観点で単語を抽出したい時、幹部クラスを選択する。Ｓ４０２～Ｓ４０３を、該当する単語カウントテーブルの全単語に対して行うまで、繰り返す。選択したクラスの単語カウントテーブルから単語を１つ選択する（Ｓ４０２）。例えば、技術クラスの単語カウントテーブル１１から、単語「画期的」を選択する。

【００３４】

次に、該当する単語カウントテーブルのカウンタ値と、他の単語カウントテーブルのカウンタ値を参照して、選択した単語の重要度を計算し、対応する単語重要度テーブルに登録する（Ｓ４０３）。この時、各クラスの文書数などを考慮に入れてもよい。

【００３５】

例えば、単語ｋの重要度Ｉを以下の式によって求めることが出来る。

【００３６】

【数１】

$$I(k) = \frac{\text{該当クラスにおけるkのカウンタの値}}{\text{該当クラスの文書数}} \quad (1)$$

$$- \frac{\text{他のクラスにおけるkのカウンタの値}}{\text{他のクラスの文書数}} \quad (2)$$

【００３７】

このように重要度の決定に差分を用いると、カウンタ値の絶対値の大きさがＩ（ｋ）の反映されることになり、より多く出現したキーワードを抽出したい場合に適している。Ｓ４０２～Ｓ４０３を、該当する単語カウントテーブルの全単語に対して行うまで、繰り返す（Ｓ４０４）。全てのクラスに対して、Ｓ４０１～Ｓ４０４を繰り返す（Ｓ４０５）。

【００３８】

結果として、各クラスに対する単語重要度テーブル１４、１５、１６を得る（図１４、図１５、図１６）。

【００３９】

図１７は、キーワード抽出手段５における請求項１記載の方法のフローチャートである。Ｓ５０１～Ｓ５０３を、全てのキーワード抽出対象のクラスに対して行うまで繰り返す。まず、キーワード抽出対象のクラスを１つ選択する（Ｓ５０１）。そのクラスに対応する単語重要度テーブルを重要度でソートし、ソートした単語重要度テーブルを作成する（Ｓ５０２）。そして、ソートした単語重要度テーブル上位の単語を抽出キーワードとする（Ｓ５０３）。Ｓ５０１～Ｓ５０３を、全てのキーワード抽出対象のクラスに対して行うまで繰り返す（Ｓ５０４）。

【００４０】

結果として、ソートした単語重要度テーブル１７、１８、１９を得る（図１８、図１９、図２０）。上述した、一連の処理を行うことで、対象の文書集合から各クラスの観点での特徴キーワードを抽出することが出来る。

【００４１】

図２１は、キーワード抽出手段５における請求項２記載の方法のフローチャートである。全てのクラスに対して、Ｓ５１１～Ｓ５１６を繰り返す。対象のクラスを１つ選択する（Ｓ５１１）。次に、選択クラスに対応する単語重要度テーブルの中の全ての単語に対して、Ｓ５１２～Ｓ５１７を繰り返す。対象の単語を１つ選択する（Ｓ５１２）。選択単語が単語統計テーブル２０に登録されているかどうかを調べ（Ｓ５１３）、登録されていれば、単語統計テーブル２０の対応する単語の統計値に、選択単語の重要度の絶対値を加え、さらに、選択クラスの単語カウントテーブルを参照し、対応する単語のカウンタに、選択単語のカウンタ値を加える（Ｓ５１４）。

【００４２】

登録されていなければ、単語統計テーブル２０に選択単語を登録した後（Ｓ５１５）、単語統計テーブル２０の対応する単語の統計値に、選択単語の重要度の絶対値を設定し、さらに、選択クラスの単語カウントテーブルを参照し、対応する単語のカウンタに、選択単語のカウンタ値を設定する（Ｓ５１６）。統計値に対しては、重要度の自乗などを加えるようにしても良い。Ｓ５１２～Ｓ５１６を

、全ての単語に対して行うまで繰り返す（Ｓ５１７）。Ｓ５１１～Ｓ５１７を、全てのクラスに対して行うまで繰り返す（Ｓ５１８）。

【００４３】

最後に、単語統計テーブル２０を統計値が小さく、カウンタ値が大きい順にソートし、ソートした単語統計テーブル２１を作成する（Ｓ５１９）。この時、具体的には、例えば

係数Ｃ　×　統計値　－　カウンタ値

の値を基準にソート、統計値とカウンタ値のどちらかに閾値を設定することで絞り込み残りの値でソート、統計値／カウンタ値の値を基準にソート、などとすれば良い。そして、ソートした単語統計テーブル２１の上位単語を、抽出キーワードとする（Ｓ５２０）。

【００４４】

結果として得られた、ソートした単語統計テーブル２１を図２２に示す。ここでは、統計値／カウンタ値の値の小さい順にソートした。上述した、一連の処理を行うことで、対象の文書集合から各クラスによって出現傾向の似ているキーワードを抽出することが出来る。

【００４５】

本実施例では、該当クラスと他のクラスとの差分に着目したが、比率に着目することも出来る。すなわち、図１３における、Ｓ４０３の重要度Ｉの計算において、

【００４６】

【数２】

$$I(k) = \frac{\text{該当クラスにおける}k\text{のカウンタの値}}{\text{該当クラスの文書数}} \quad (3)$$

$$\div \frac{\text{他のクラスにおける}k\text{のカウンタの値}}{\text{他のクラスの文書数}} \quad (4)$$

【００４７】

とすればよい。このように重要度の決定に比率を用いると、文書数の絶対数に関わらず重要度の大小の幅が大きくなり、文書数によらずに重要度を考慮したキー

ワードを抽出したい場合に適している。

【００４８】

この時、統計値の計算をする時に、無限大なる場合を考慮し、重要度に上限、下限を設定し、さらに、重要度が１未満ならその逆数を重要度の代わりに使うようにするか、対数の絶対値を使うなど、重要度の違いが累積されるようにする。または、重要度が１未満の時その逆数を重要度の代わりに使う場合には、重要度またはその逆数が１以上になることに注目し、統計値に加えるのではなく、乗じることも出来る。

【００４９】

また、本実施例では、対象の文書集合を開発資料とし、分割基準を著者としたが、その他分割基準になりうる所望の基準で良く、例えば、メールを対象として、社外用のメールと、社内用のメールに分割したり、メールの発生月日または時刻で分割したり、特定の相手とそれ以外に分割するなど、対象データに付加されているものであれば何でも良い。

【００５０】

様々な検索システムでは、利用者のアクセスをログとして記録している場合が多い。この時、アクセスログにはアクセスした利用者のＩＤや、アクセス日時や、検索キー（多くの場合キーワード）が記録されている。従って、この一回のアクセスに対するログを１文書として扱うと、アクセスログ全体で、文書集合を構成することになり、利用者ＩＤで分割したり、アクセス日時を用いて昼夜あるいは平日、休日などに分割することで、それぞれのアクセス傾向を調べることが出来る。

【００５１】

【発明の効果】

本発明のキーワード抽出方法、キーワード抽出装置、又は記録媒体を用いることによれば、種々のパラメータで分割可能な大量のデータに対して、分割されたグループ毎の統計処理結果の違いを比較することによって、分割されたグループ毎、又は全データのキーワードを抽出し、全データの特徴と、全データの中での分割されたグループの特異性及び／又は傾向を把握することが出来、しかも、デ

ータの形式として特定のものは必要でなく、所望のものを扱える。

【００５２】

特に、重要度の決定に、出現頻度の差分を用いる方法では、より出現頻度の高いキーワードを抽出することが出来、重要度の決定に、出現頻度の比率を用いる方法では、文書の絶対数によらずに重要度を考慮したキーワードを抽出することが出来る。

【図面の簡単な説明】

【図１】

本発明の機能ブロック図である。

【図２】

データ対象選択手段における動作のフローチャートである。

【図３】

文書テーブルの例を示す説明図である。

【図４】

対象データ分割手段における動作のフローチャートである。

【図５】

幹部クラスの単語リストの例を示す説明図である。

【図６】

企画クラスの単語リストの例を示す説明図である。

【図７】

技術クラスの単語リストの例を示す説明図である。

【図８】

著者クラステーブルの例を示す説明図である。

【図９】

部分統計処理手段における動作のフローチャートである。

【図１０】

幹部クラスの単語カウントテーブルの例を示す説明図である。

【図１１】

企画クラスの単語カウントテーブルの例を示す説明図である。

【図１２】

技術クラスの単語カウントテーブルの例を示す説明図である。

【図１３】

部分統計処理結果比較手段における動作のフローチャートである。

【図１４】

幹部クラスの単語重要度テーブルの例を示す説明図である。

【図１５】

企画クラスの単語重要度テーブルの例を示す説明図である。

【図１６】

技術クラスの単語重要度テーブルの例を示す説明図である。

【図１７】

キーワード抽出手段における請求項１記載の方法のフローチャートである。

【図１８】

幹部クラスのソートした単語重要度テーブルの例を示す説明図である。

【図１９】

企画クラスのソートした単語重要度テーブルの例を示す説明図である。

【図２０】

技術クラスのソートした単語重要度テーブルの例を示す説明図である。

【図２１】

キーワード抽出手段における請求項２記載の方法のフローチャートである。

【図２２】

単語統計テーブルの例を示す説明図である。

【図２３】

ソートした単語統計テーブルの例を示す説明図である。

【符号の説明】

1　対象データ選択手段

2　対象データ分割手段

3　部分統計処理手段

4　部分統計処理結果比較手段

５　キーワード抽出手段

６　文書テーブル

７　幹部クラスの単語リスト

８　企画クラスの単語リスト

９　技術クラスの単語リスト

１０　著者クラステーブル

１１　幹部クラスの単語カウントテーブル

１２　企画クラスの単語カウントテーブル

１３　技術クラスの単語カウントテーブル

１４　幹部クラスの単語重要度テーブル

１５　企画クラスの単語重要度テーブル

１６　技術クラスの単語重要度テーブル

１７　幹部クラスのソートした単語重要度テーブル

１８　企画クラスのソートした単語重要度テーブル

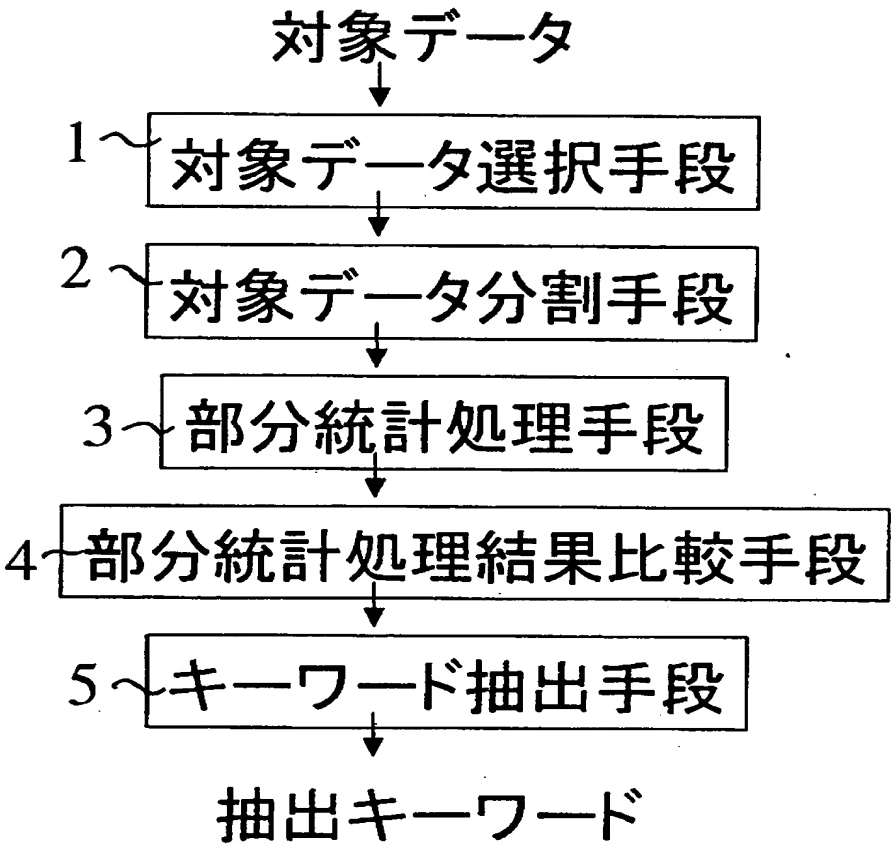
１９　技術クラスのソートした単語重要度テーブル

２０　単語統計テーブル

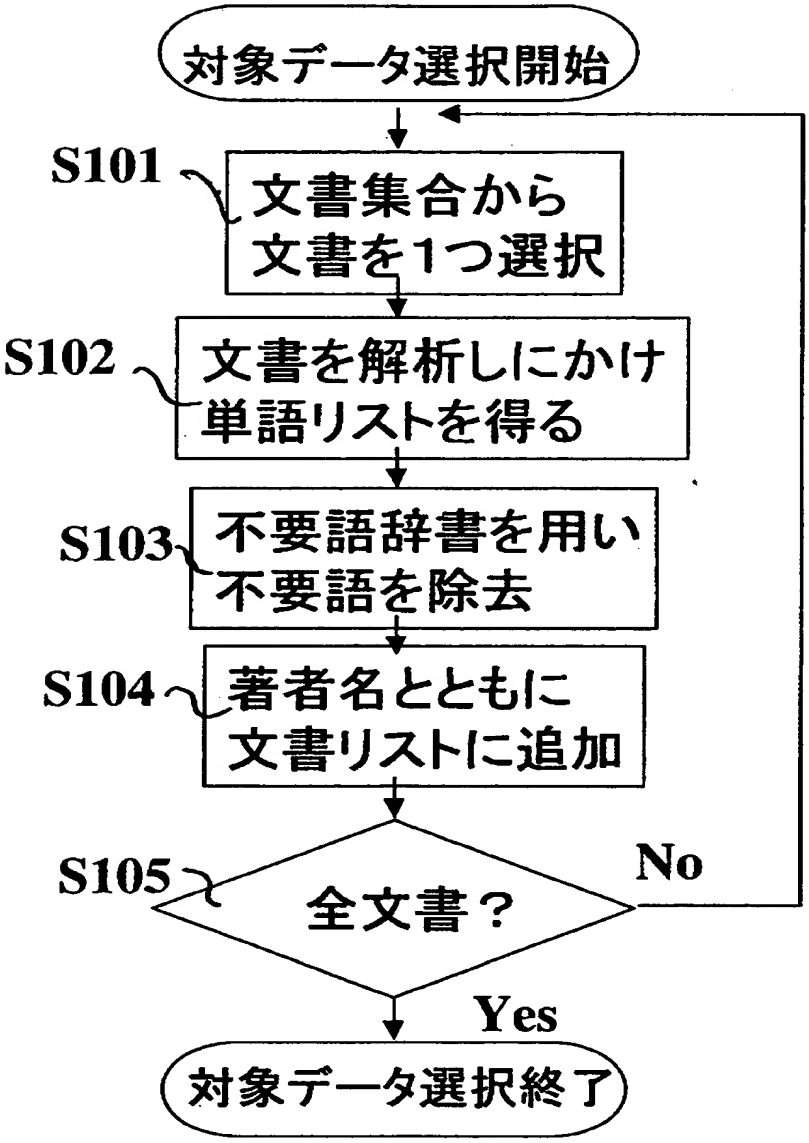
２１　ソートした単語統計テーブル

【書類名】　図面

【図１】


対象データ
1
対象データ選択手段
2
対象データ分割手段
3
部分統計処理手段
4
部分統計処理結果比較手段
5
キーワード抽出手段
抽出キーワード

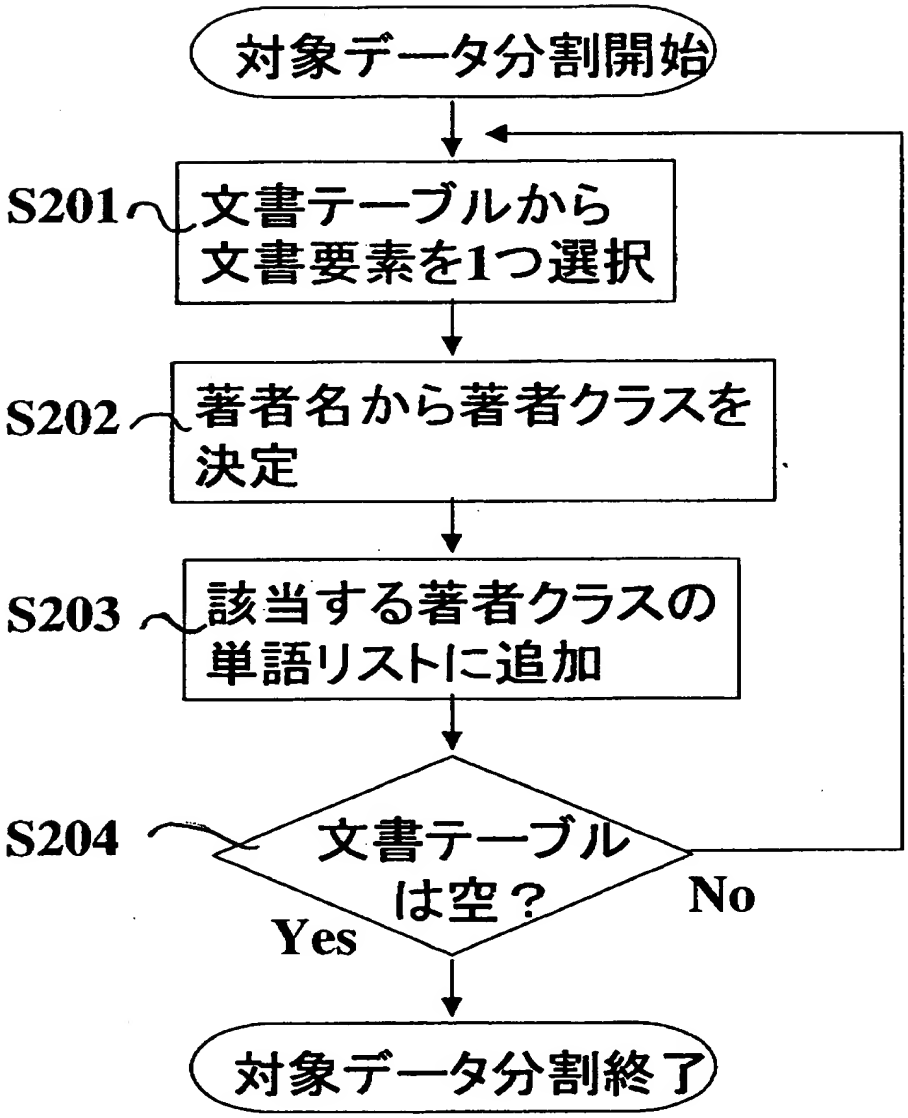
【図２】


対象データ選択開始
S101
文書集合から
文書を1つ選択
S102
文書を解析しにかけ
単語リストを得る
S103
不要語辞書を用い
不要語を除去
S104
著者名とともに
文書リストに追加
S105
全文書？
No
Yes
対象データ選択終了

【図３】

文書番号	著者	単語(出現数)
1	幹部A	画期的(10), 技術革新(5), デジタル(4)
2	企画A	多彩(7), デジタル(3), 高画質(2), 画期的(2)
3	技術A	デジタル(6), 通信(5), 詳細仕様(3)
4	技術B	デジタル(4), 通信(8)
5	幹部B	画期的(8), デジタル(6), 技術革新(7), 情報化社会(5), 高画質(4)
6	技術C	技術革新(3), 高画質(6)
7	幹部B	技術革新(6), 経営環境(1), 情報化社会(5)
8	企画B	多彩(3), 高画質(3), 鮮やか(4), 通信(2)
9	幹部C	経営環境(5), 通信(2), デジタル(8)
10	幹部C	画期的(7), 技術革新(4)
11	技術C	詳細仕様(3), デジタル(8), プロトコル(5)
12	企画C	多彩(6), デジタル(4), 鮮やか(4)

【図４】


対象データ分割開始
S201
文書テーブルから
文書要素を1つ選択
S202
著者名から著者クラスを
決定
S203
該当する著者クラスの
単語リストに追加
S204
文書テーブル
は空？
No
Yes
対象データ分割終了

【図５】

単語
画期的(10), 技術革新(5), デジタル(4),
画期的(8), デジタル(6), 技術革新(7), 情報化社会(5), 高画質(4),
技術革新(6), 経営環境(4), 情報化社会(5),
経営環境(5), 通信(2), デジタル(8),
画期的(7), 技術革新(4)

【図６】

単語
多彩(7), デジタル(3), 高画質(2), 画期的(2),
多彩(3), 高画質(3), 鮮やか(4), 通信(2),
多彩(6), デジタル(4), 鮮やか(4)

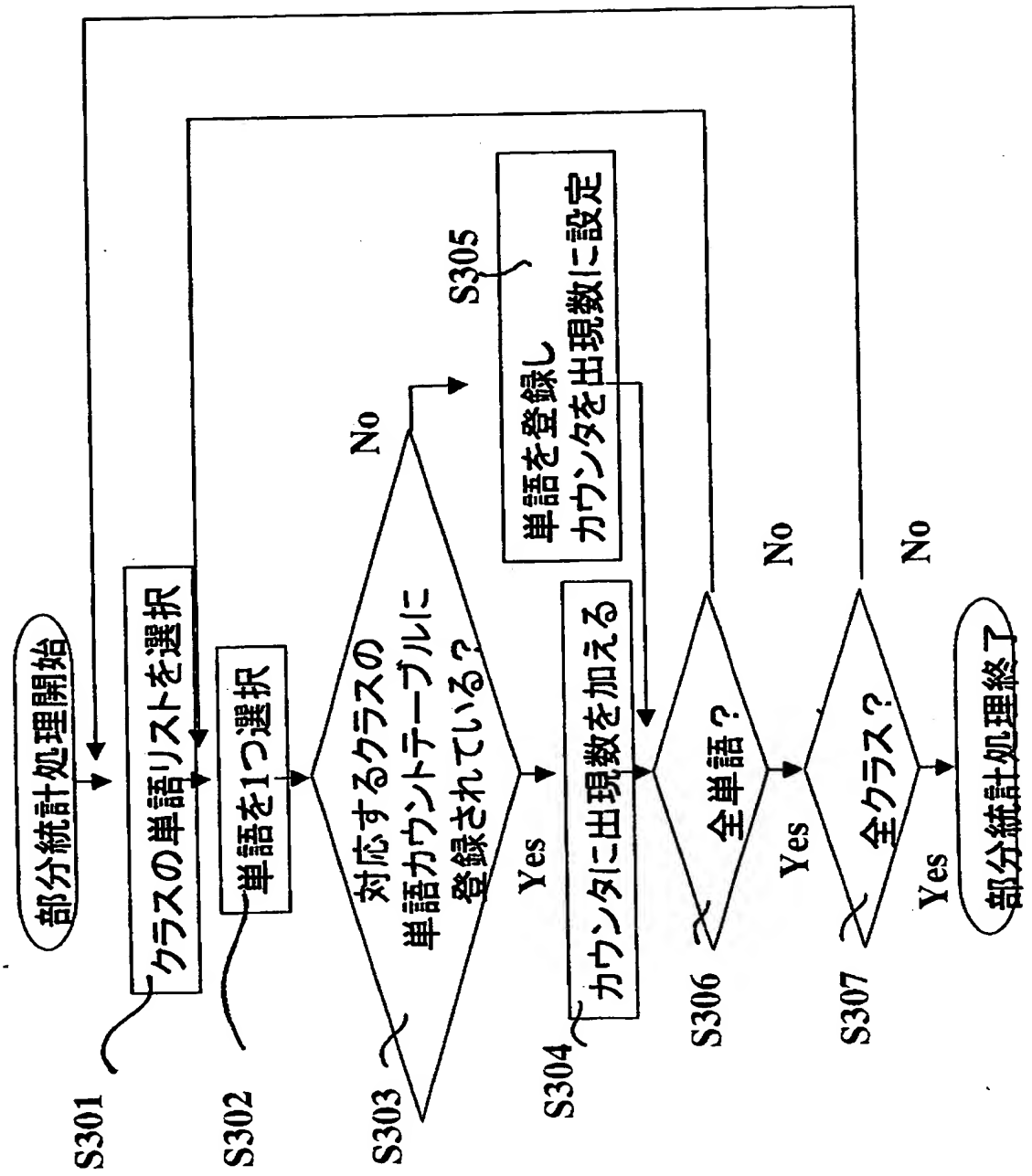
【図７】

単語
デジタル(6), 通信(5), 詳細仕様(3),
デジタル(4), 通信(8),
技術革新(3), 高画質(6),
詳細仕様(3), デジタル(8), プロトコル(5)

【図８】

著者名	著者クラス
幹部A	幹部
幹部B	幹部
幹部C	幹部
企画A	企画
企画B	企画
企画C	企画
技術A	技術
技術B	技術
技術C	技術

【図９】


部分統計処理開始
クラスの単語リストを選択
S301
単語を1つ選択
S302
対応するクラスの
単語カウントテーブルに
登録されている？
S303
No
Yes
S305
単語を登録し
カウンタを出現数に設定
カウンタに出現数を加える
S304
全単語？
S306
No
Yes
全クラス？
S307
No
Yes
部分統計処理終了

【図１０】

単語	カウンタ
画期的	25
技術革新	22
デジタル	18
情報化社会	10
経営環境	9
高画質	4
通信	2

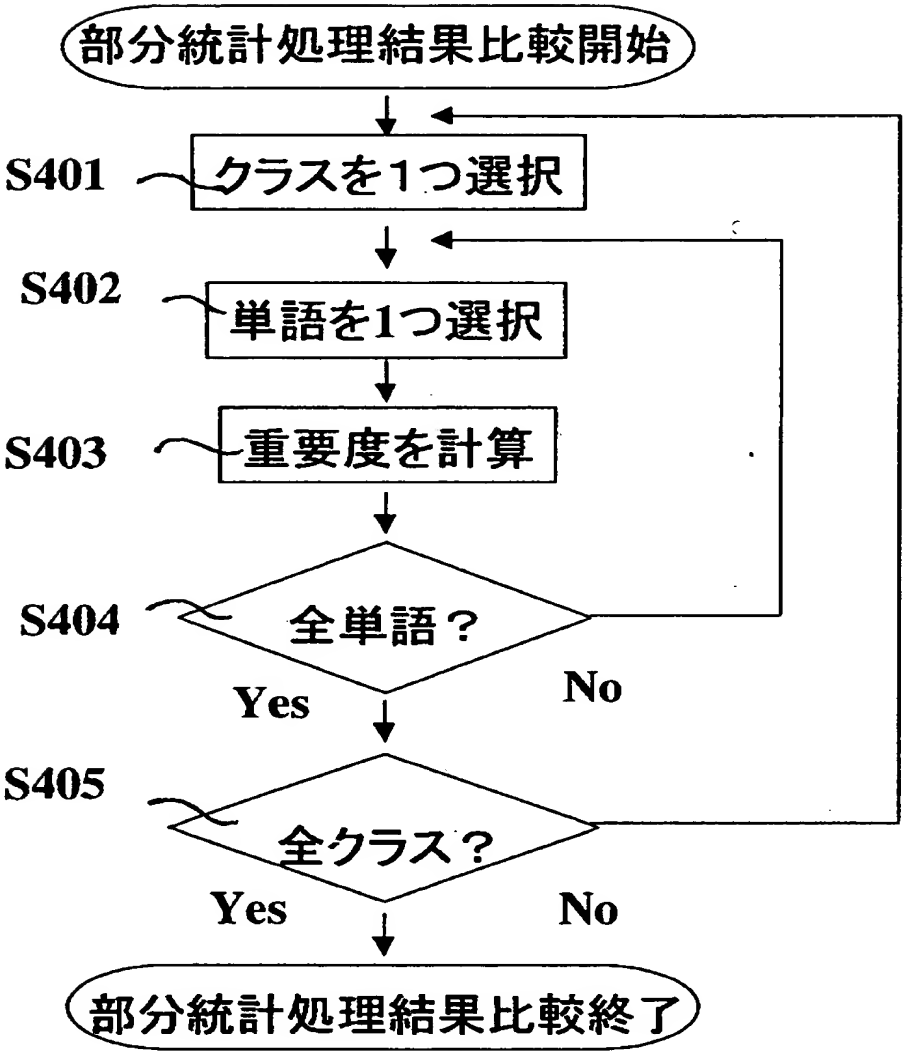
【図１１】

単語	カウンタ
多彩	16
鮮やか	8
デジタル	7
高画質	5
画期的	2
通信	2

【図１２】

単語	カウンタ
デジタル	18
通信	13
詳細仕様	6
技術革新	3
高画質	6
プロトコル	5

【図１３】


部分統計処理結果比較開始
S401
クラスを1つ選択
S402
単語を1つ選択
S403
重要度を計算
S404
全単語？
Yes
No
S405
全クラス？
Yes
No
部分統計処理結果比較終了

【図１４】

単語	重要度
画期的	4.714
技術革新	3.971
経営環境	1.800
デジタル	0.028
情報化社会	2.000
高画質	-0.771
通信	-1.742

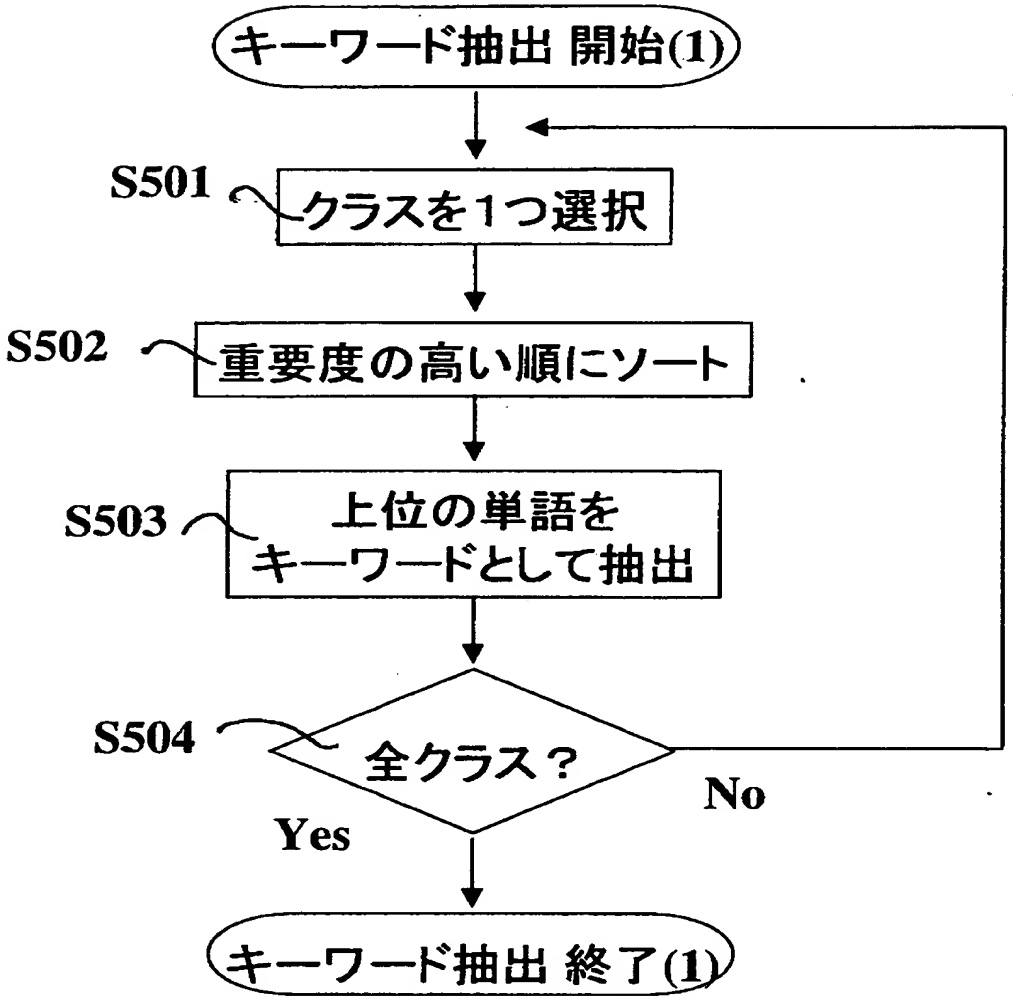
【図１５】

単語	重要度
多彩	5.333
高画質	0.555
鮮やか	2.666
デジタル	-1.666
画期的	-2.111
通信	-1.000

【図１６】

単語	重要度
デジタル	1.375
詳細化様	1.500
通信	2.750
技術革新	-2.000
高画質	0.375
プロトコル	1.250

【図１７】



出証特平１１－３０５３９２４

【図１８】

単語	重要度
画期的	4.714
技術革新	3.971
情報化社会	2.000
経営環境	1.800
デジタル	0.028
高画質	-0.771
通信	-1.742

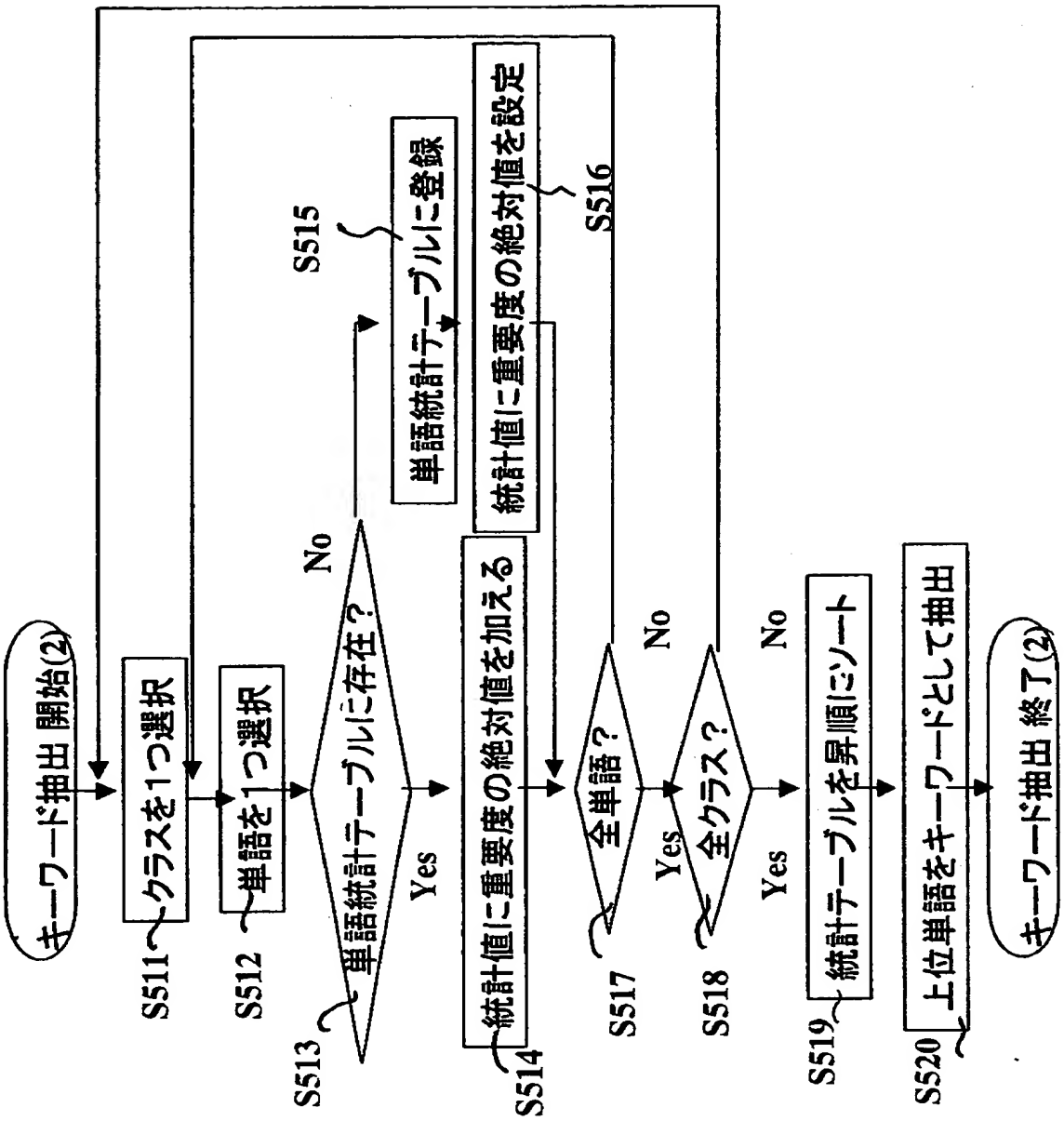
【図１９】

単語	重要度
多彩	5.333
鮮やか	2.666
高画質	0.555
通信	-1.000
デジタル	-1.666
画期的	-2.111

【図２０】

単語	重要度
通信	2.750
デジタル	1.375
詳細仕様	1.500
プロトコル	1.250
高画質	0.375
技術革新	-2.000

【図２１】


キーワード抽出 開始(2)
クラスを1つ選択
S511
単語を1つ選択
S512
単語統計テーブルに存在？
S513
No
Yes
S515
単語統計テーブルに登録
統計値に重要度の絶対値を加える
S514
統計値に重要度の絶対値を設定
S516
全単語？
S517
No
Yes
全クラス？
S518
No
Yes
統計テーブルを昇順にソート
S519
上位単語をキーワードとして抽出
S520
キーワード抽出 終了(2)

【図２２】

単語	統計値	カウンタ
画期的	6.825	27
技術革新	5.971	25
情報化社会	2.000	10
経営環境	1.800	9
デジタル	3.069	43
高画質	1.701	15
通信	5.492	17
詳細仕様	1.500	6
多彩	5.333	16
鮮やか	2.666	8
プロトコル	1.250	5

【図２３】

単語	統計値	カウンタ	統計値/カウンタ
デジタル	3.069	43	0.071
高画質	1.701	15	0.113
情報化社会	2.000	10	0.200
経営環境	1.800	9	0.200
技術革新	5.971	25	0.238
詳細仕様	1.500	6	0.250
プロトコル	1.250	5	0.250
画期的	6.825	27	0.252
通信	5.492	17	0.323
多彩	5.333	16	0.333
鮮やか	2.666	8	0.333

【書類名】　　要約書

【要約】

【課題】　キーワードを抽出したいデータの形状に依存せず、データを所望のパラメータで分割し、キーワードを抽出する。

【解決手段】　データからキーワードを抽出する方法であって、所望のパラメータを用いて前記データを分割するステップと、分割されたグループ毎に単語を統計処理するステップと、統計処理された結果を比較し、重要度を算出するステップと、算出された重要度から、比較的重要度が高いと判断された単語からキーワードを決定するステップと、を備えたことを特徴とする。

【選択図】　なし

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　000005049
【住所又は居所】　大阪府大阪市阿倍野区長池町２２番２２号
【氏名又は名称】　シャープ株式会社
【代理人】　申請人
【識別番号】　100103296
【住所又は居所】　大阪府大阪市阿倍野区長池町２２番２２号　シャープ株式会社内
【氏名又は名称】　小池　隆彌

出 願 人 履 歴 情 報

識別番号　〔０００００５０４９〕

１．変更年月日　１９９０年　８月２９日
［変更理由］　新規登録
住　所　大阪府大阪市阿倍野区長池町２２番２２号
氏　名　シャープ株式会社